D990–D995 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 05 November 2014*
doi: 10.1093/nar/gku1070

Plastid-LCGbase: a collection of evolutionarily conserved plastid-associated gene pairs

Dapeng Wang[1,2,*] and Jun Yu[1,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, P. R. China and [2]Stem Cell Laboratory, UCL Cancer Institute, University College London, London WC1E 6BT, UK

Received August 14, 2014; Revised October 13, 2014; Accepted October 16, 2014

ABSTRACT

Plastids carry their own genetic material that encodes a variable set of genes that are limited in number but functionally important. Aside from orthology, the lineage-specific order and orientation of these genes are also relevant. Here, we develop a database, Plastid-LCGbase (http://lcgbase.big.ac.cn/plastid-LCGbase/), which focuses on organizational variability of plastid genes and genomes from diverse taxonomic groups. The current Plastid-LCGbase contains information from 470 plastid genomes and exhibits several unique features. First, through a genome-overview page generated from OrganellarGenomeDRAW, it displays general arrangement of all plastid genes (circular or linear). Second, it shows patterns and modes of all paired plastid genes and their physical distances across user-defined lineages, which are facilitated by a stepwise stratification of taxonomic groups. Third, it divides the paired genes into three categories (co-directionally-paired genes or CDPGs, convergently-paired genes or CPGs and divergently-paired genes or DPGs) and three patterns (separation, overlap and inclusion) and provides basic statistics for each species. Fourth, the gene pairing scheme is expandable, where neighboring genes can also be included in species-/lineage-specific comparisons. We hope that Plastid-LCGbase facilitates gene variation (insertion-deletion, translocation and rearrangement) and transcription-level studies of plastid genomes.

INTRODUCTION

Plastid is a vital organelle for the photosynthesis of eukaryotic species in broad taxa, including plants, alga and protists; it is also regarded as a favorable genetic material for transformation and manipulation for its extra-chromosomal status (1,2). The structure and number of plastids are different from those of mitochondrion and other subcellular organelles, as well as their phenotypes that are influenced by not only genetics but also environmental factors (3,4). According to the endosymbiotic theory, plastid originated from a cyanobacterium and experienced multiple evolutionary events, which had altered their primary, secondary and tertiary structures (5) to the extent that not every plastid contains a typical genome and not all plastid genes are involved in photosynthesis (6–8). The basic structure of plastid genomes, containing a number of essential protein-coding genes as well as rRNAs and tRNAs, is divided into four parts: a long single-copy, a small single-copy and two inverted repeated segments (9–13). The gene flow among plastid, mitochondrial and nuclear genomes starts at very early stages but may have heterogeneous rates (14–16). Plastid genomes produce considerable amount of essential and indispensable functional proteins for various functions, such as photosynthesis, respiration and translation (17), and the rest participants, proteins and RNAs, are contributed by nuclear genes, of which some are proposed to originate from cyanobacterium (18). The number and type of plastid genes vary among species and lineages but as a functional set still adequately maintains essential functions (17). In general, eukaryotic genomes are organized into gene clusters and these clustered genes often collaborate for function and appear not acting alone (19–22). For plastid genomes, it is known that there are several operon-like gene clusters that are co-regulated or co-expressed, composed of neighboring or consecutively ordered genes, and lead to improvement of transcription and translation efficiency (23).

Most of the current plastid relevant databases, such as GOBASE (24) and ChloroplastDB (25), emphasize gene structure and sequence annotation but pay less attention to genome organization. The only study related to plastid gene order included 32 species and merely displayed the text information (26), although there have been several databases built for visualizing the nuclear gene order

*To whom correspondence should be addressed. Tel: +86 10 84097898; Fax: +86 10 84097540; Email: junyu@big.ac.cn
Correspondence may also be addressed to Dapeng Wang. Tel: +44 20 76797602; Fax: +44 20 76796817; Email: dapeng.wang@ucl.ac.uk

in different evolutionary scales or in a limited scope of taxonomic grouping (27–32). As the rising popularity of high-throughput sequencing and the rapid accumulation of organellar genome sequences (33), the public data collection now has 470 plastid genomes. Here, based on conserved paired genes, our plastid-LCGbase provides general survey, visualization, comparative frameworks for plastid genomes in various user-defined phylogenetic grouping. We also define six patterns or modes and eight types of transcription start sites (TSS) distances for the plastid gene pairs. This database should become a useful repository for the study of plastid genome alignment and arrangement study as well as for the discovery of possible co-regulation of adjacent genes over evolutionary time scales.

MATERIALS AND METHODS

We collected genome information and sequences of protein-coding genes and non-coding RNAs of 470 plastids from NCBI Organelle Genome Resources (web and ftp server) based on careful selection of representative species. Taxonomic data were downloaded from the NCBI taxonomy ftp site and the keywords include names of kingdom, phylum, class, order, family, genus and species. The circular or linear maps for plastid genomes were drawn by using OrganellarGenomeDRAW (v1.1.1) (34) and other figures were plotted by using R software package. Comparisons between a reference genome and other genomes (99 in number) that are most similar to the reference were manipulated by using CGView (35). Similar determination was carried out by using Blastp (ncbi-blast-2.2.28+) with E-value = 1e-5 and max_target_seqs = 500 (36) for protein-coding genes and BlastR with E-value = 1e-5 for non-coding RNAs (37). Classification of gene families was based on TribeMCL (Markov Clustering) with I = 2 (38). The definition of conserved gene pairs was performed in a visual system. In details, once core data sets are imported into MYSQL database, an optimal index is created to make sure for fast user inquiry. PHP takes charge of the calculation modules, makes the reference chromosome fixed and searches along two opposite directions in other chromosomes in the process of comparisons. Once the computation is finished, a figure containing gene arrangement in both the reference and the searchable chromosomes is shown as an image, together with three types of textual formats. Colors and arrows are used to indicate homologous groups and transcriptional orientations, respectively; the colors are forced to distinguish different homologs. We recommend using 'Google Chrome Browser' to view this database since we have tested that this browser works compatible with different operating system for the full database functions.

RESULTS

We started by constructing phylogenetic trees for species involved using CVTree (39) based on proteomic data to provide a glance of evolutionary relationship among all the species for users (Figure 1A). In general, the database offers a genome map function to show an overview of gene distributions on the browse page (Figure 1G). At the same time, it provides graphic views for structural changes among plastid genomes at a global level, which include both DNAs and translated coding sequences (CDSs). For better display of structural features, we divided plastid genome into three gene groups: protein-coding, non-coding RNA and all genes (including the previous two groups). One of the functions for the database is to define paired genes into all three types and to discover conserved patterns of the gene pairs in different evolutionary lineages. In the search page, we provide eight different colors to distinguish the distance of neighboring genes (0–300 bp, 300–500 bp, 500–800 bp, 800–1000 bp, 1000–1200 bp, 1200–1500 bp, 1500–2000 bp and > 2000 bp) and multiple-checked boxes to determine the species of interest on the sorted display of taxa from kingdom all the way down to genus (Figure 1B). When a gene identifier is entered by a user, the resulting page produces a figure containing a list of conserved gene pairs (both homology-based and strand-specific) (Figure 1C). Since the distances between paired genes are color-coded, the dynamics of TSS of homologous gene pairs in different species can be visually compared. If query gene is unknown, the database provides two alternative choices since all featured data have been summarized in the species table (Figure 1D). One way is to browse the gene list in particular genomes to find their names in various nomenclature system (e.g. Gene Identifier, Protein ID, Gene ID and Product) and position information (e.g. Strand, Start and End) (Figure 1E). Another way is to view the gene pair list including their relationship and individual features (Figure 1F). We also calculated all conserved gene pairs in the 470 plastid genomes for browsing and downloading. Furthermore, we define operon-like structures as determined by concatenating highly-conserved gene pairs (at least conserved in 100 plastid genomes) in certain species. In addition, we classify gene pairs into nine categories based on whether they are co-directionally-paired genes (CDPGs), convergently-paired genes (CPGs) or divergently-paired genes (DPGs) and in 'Separation', 'Overlap' and 'Inclusion' as patterns. The former is an orientation parameter that defines gene clusters based on relative transcription direction of neighboring genes; the latter is a distance parameter that characterizes physical distance of neighboring genes (Figure 2). In addition, we plot densities of TSS distance in logarithmic scale for CDPGs, CPGs, DPGs (Figure 1H) and all paired genes, and show barplots of all nine paired gene types on the 'Parameter' page (Figure 1I). We offer processed gene pair data of all plastid genomes for free-download by users. Every figure in this database can be enlarged to display a high-resolution version. In order to establish connections between this database and external public databases, we linked many keywords to their NCBI definitions and annotation pages; for example, 'Species', 'Protein GI', 'Locus', 'Protein Accession' and 'Gene ID' are all appropriately linked.

DATA OVERVIEW

In the database, we classified 470 plastids into 9 categories (*Alveolata*, *Cryptophyta*, *Euglenozoa*, *Glaucocystophyceae*, *Haptophyceae*, *Rhizaria*, *Rhodophyta*, *Stramenopiles* and *Viridiplantae*), 111 orders and 152 families, albeit some incomplete information for their order and family definitions. Most genomes are circular except 10 linear displays. We

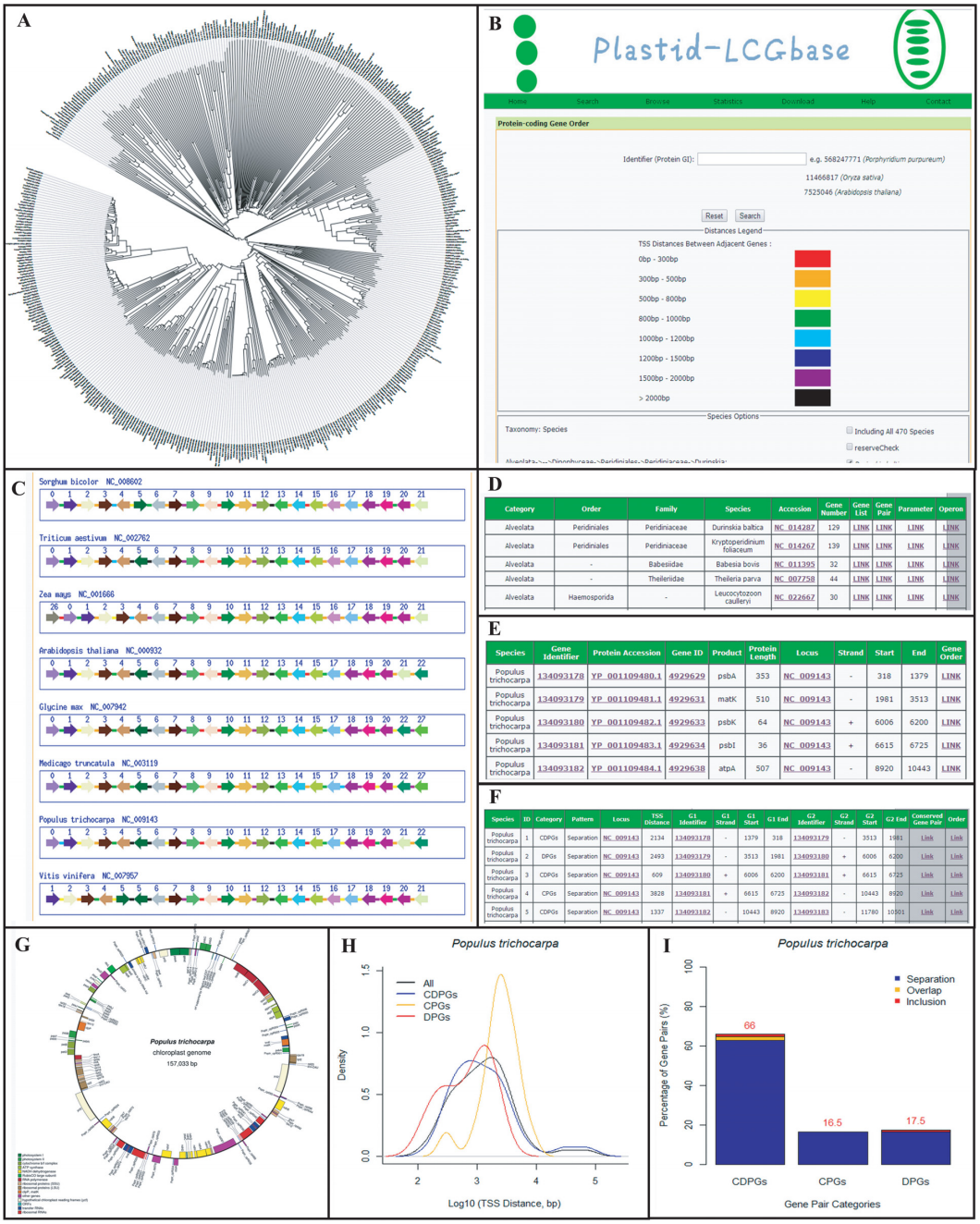

Figure 1. Screenshots for major functional modules of Plastid-LCGbase. (**A**) Phylogenetic trees for 470 species built from whole plastid proteomes. (**B**) The search page for determining conserved gene pairs and visualizing gene orders. (**C**) The result page from search pages. Arrows and orientation indicate genes and their transcription direction. (**D**) The browse page. (**E**) The gene list page. (**F**) The gene pair page. (**G**) The genome map page. (**H**) Distributions of TSS distances from the three types of gene pairs. (**I**) Barplots of the three types and three patterns of gene pairs.

adopted four inflation parameters (I = 1.4, I = 2, I = 3 and I = 4) to deduce gene family classification for all the proteomes and found that the shape of their distributional curves are quite similar (data not shown). I = 1.4 generates more large gene families while I = 4 leads to more small gene families. We decided to choose a moderate one (I = 2) for the analysis. In details, the largest gene family contains 911 members for protein-coding genes and 857 members for non-coding RNAs; other measurements for gene family sizes are: 62 families for protein-coding genes and 25 families for non-coding RNAs > 400 members; 80 families and 36 families > = 100 members and 137 families and 47 families > = 30 members. We calculated some parameters for each genome to observe the complexity and sampled some representatives (Table 1). First, there are cases with extremely properties, such as the plastid of the parasitic *Babesia bovis*' (category: *Alveolata*; family: *Babesiidae*); it is much smaller than the median value

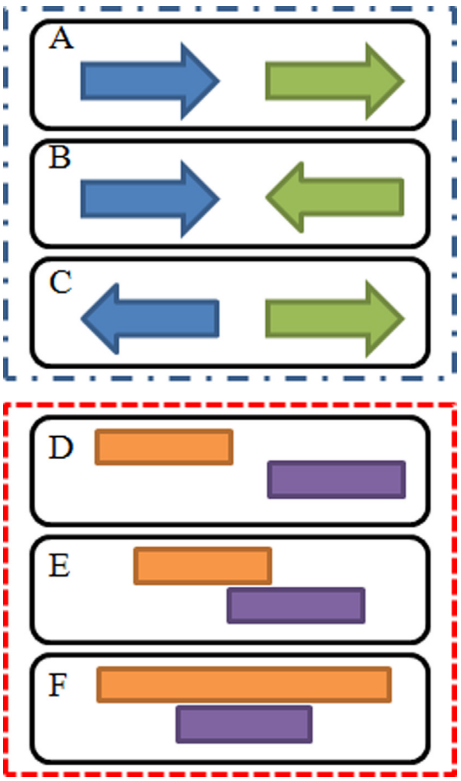

Figure 2. The definition of categories and patterns of gene pairs. On the basis of relative transcription direction, gene pairs are divided into three categories: (**A**) CDPG, gene transcribed in the same direction. (**B**) CPG, gene transcribed in the opposing direction but toward each other. (**C**) DPG, gene transcribed in the opposing transcription direction but toward distant directions. According to the relative positions of neighboring genes, gene pairs are classified into three patterns: (**D**) 'Separation', no overlapping regions. (**E**) 'Overlap', shared some regions but no one covers the other completely. (**F**) 'Inclusion', one gene covers all regions of the other gene.

and has the strongest strand imbalance (i.e, all genes on the same strand). The plastid of *Porphyra purpurea*' (category: *Rhodophyta*; family: *Bangiaceae*) is another interesting example, whose genome is much larger and has more DPGs. Second, we computed strand ratio to estimate the strand biased gene distribution and found that the median value is larger than 1, which indicates that the strand bias is common among plastid genomes. Third, CDPGs are the most popular patterns and the percentage of CPGs and DPGs are often comparable. The median percentages of 'Separation', 'Overlap' and 'Inclusion' are 95.7%, 3.4% and 1.2%, respectively (data not shown), which suggests that most of the gene pairs are not overlapping. Last, the median length of transcripts is smaller than the median distance between TSS, despite that their extents vary among different species. It is also noted that the current genomic data sets cover a large scope of robust data structures.

CASE STUDIES

The first case

A pair of CDPG, *atpG* (568247771) and *atpF* (568247772), responsible for ATP synthesis, from *Porphyridium purpureum* (Category: *Rhodophyta*; Order: *Porphyridiales*; Family: *Porphyridiaceae*) and their counterparts in other species appear in several lower plant species (*Alveolata*, *Cryptophyta*, *Glaucocystophyceae*, *Haptophyceae*, *Rhodophyta* and *Stramenopiles*) but they are absent in all *Viridiplantae*. This observation indicates that this gene pair has an ancient origin but suffered from gene loss when species are evolving. The TSS distances between the gene pair range 500–800 bp in most species with an exception of *Porphyridium purpureum*, whose TSS distance is slightly larger: around 1000–1200 bp. Furthermore, a conservation pattern appears expanded to their four neighboring genes, forming a cluster of *atpI-atpH-atpG-atpF-atpD-atpA* in all species but not in *Cyanophora paradoxa*, due to the loss of *atpI*, reflecting the unique feature of *Glaucocystophyceae* (Supplementary Figure S1).

The second case

A pair DPG involving *psbN* (11466817) and *psbH* (11466818), are part of the photosystem II in *Oryza sativa* (Category: *Viridiplantae*; Order: *Poales*; Family: *Poaceae*), which is shared among 432 species. Their TSS distances are very small (0–300 bp) in most species but become larger in *Nephroselmis olivacea* (300–500 bp), *Chlamydomonas reinhardtii* (500–800 bp) and *Pleodorina starrii* (500–800 bp). The short TSS is ancient pattern since it exists in both *Viridiplantae* and non-*Viridiplantae* species. When looking at the Family *Fabaceae*, the gene clusters containing the pair becomes separated in different species, showing subtle differences. For example, a cluster of 20 consecutive genes concerning the pair (*petL-petG-psaJ-rpl33-rps18-rpl20-rps12-clpP-psbB-psbT-psbN-psbH-petB-petD-rpoA-rps11-rpl36-rps8-rpl14-rpl16-rps3-rps19*) are conserved not only in nine *Glycine* subspecies (*Glycine canescens*, *Glycine cyrtoloba*, *Glycine dolichocarpa*, *Glycine falcata*, *Glycine max*, *Glycine soja*, *Glycine stenophita*, *Glycine syndetika* and *Glycine tomentella*) but also in other related family members such as *Lotus japonicus*, *Lupinus luteus*, *Medicago truncatula*, *Millettia pinnata* and *Castanea mollissima*. However, subtle changes are found in other *Fabaceae* species. When comparing *Lathyrus sativus* with *Lotus japonicus*, we observed gene inversion and insertion-deletion: the left of *psbB*, a unit of eight genes (*petL-petG-psaJ-rpl33-rps18-rpl20-rps12-clpP*), was inversed and then *rps12* was deleted between *clpP* and *rpl20* in *Lathyrus sativus* (Supplementary Figure S2).

The third case

A pair of CPGs, *petA* (7525046) and *psbJ* (7525047), of *Arabidopsis thaliana* (Category: *Viridiplantae*; Order: *Brassicales*; Family: *Brassicaceae*), is part of the cytochrome complex and photosystem II reaction center protein, respectively. Their orthologs have been identified only in *Viridiplantae*, especially in *Amborella trichopoda*, which separates from other flowering plants in the very early stage of evolution (40). In particular, most of the TSS distances of such a pair are larger than 1500 bp. Together with the observations in various species, we speculate that this cluster (*ndhJ-ndhK-ndhC-atpE-atpB-rbcL-accD-psaI-ycf4-cemA-petA-psbJ-psbL-psbF-psbE-petL-petG-psaJ-rpl33-rps18-rpl20-rps12*) is ancestral among Angiosperms.

Table 1. A basic survey for protein-coding genes of 13 species as examples of Plastid-LCGbase

Species	Genome length (nt)	Gene number	Strand ratio	CDPGs%	CPGs%	DPGs%	Median transcript length	Median TSS distance
Durinskia baltica	116470	129	1.67	79.7	10.2	10.2	419	533
Babesia bovis	35107	32	33.00	100.0	0.0	0.0	581	592
Cryptomonas paramecium	77717	82	2.11	75.3	12.3	12.3	486.5	608
Emiliania huxleyi	105309	119	1.95	73.7	13.6	12.7	416	587
Porphyra purpurea	191028	209	1.45	66.8	16.8	16.3	518	580
Cuscuta exaltata	125373	67	1.65	71.2	13.6	15.2	554	1405
Colocasia esculenta	162424	86	1.84	72.9	12.9	14.1	546.5	1230
Acidosasa purpurea	139697	82	1.21	79.0	9.9	11.1	510.5	1040
Cathaya argyrophylla	107122	70	1.32	73.9	13.0	13.0	416	1143
Aethionema cordifolium	154168	84	1.77	72.3	13.3	14.5	630.5	1009
Cicer arietinum	125319	75	1.41	74.3	12.2	13.5	605	1125
Gossypium anomalum	159507	86	1.84	75.3	11.8	12.9	579.5	1221
Allosyncarpia ternata	159593	85	1.72	70.2	14.3	15.5	605	1158.5
Median of 470 genomes	154425.5	85	1.69	74.7	12.0	13.1	554	1087.5

Note: Genome length, the length of whole genome; Gene number, the number of protein-coding genes; Strand ratio, (the number of genes in dominate strand +1)/(the number of genes in the other strand +1); CDPGs%, CPGs% and DPGs% indicate the percentages of CDPGs, CPGs and DPGs among all gene pairs. Median transcript length and median TSS distance indicate the median values of transcript length and the distance between neighboring transcription start sites.

However, there are still modifications of the cluster, which are found in different branches of plant taxa. For instance, an insertion of a hypothetical protein (134093208) between *rbcL* and *accD* in *Populus trichocarpa* and a deletion of *accD* between *rbcL* and *psaI* in several species, such as *Brachypodium distachyon* and *Triticum aestivum*, have been found (Supplementary Figure S3).

DISCUSSION AND CONCLUSION

Genomes and their genes, large or small, are always organized in order and orientation. The variation and conservation of such organizations in the context of lineages and closely related taxa and under mutation and selection over time are considered as an important part of genomic signatures. Information on plastid genomes is therefore of importance and worthy of a dedicated database. We started with analysis of paired genes to provide a window for gene co-regulation. The dynamics of neighboring gene pairs can be defined as loss of genes or loss of relationship, and it is useful in recognizing important evolutionary events and common ancestors. In fact, whole *plastome* has been used to construct phylogenetic trees for plants and to delineate the timing of speciation based on both sequence feature and gene order (41–43). We also believe that visualization of comparative genomics data helps the discovery of rules and patterns in gene orders and orientations. In addition, the precise measure of TSS distances between paired genes and the display of these distances are all useful in defining gene co-regulations, and the dynamic process of gene losses in plastid genomes and plastid-associated nuclear genes are all relevant in defining the functional network of plastid genes (44). We anticipate that plastid-LCGbase will be developed to become a principle bioinformatic resource for plastid study.

FUTURE PLANS

We have plans in mind to improve the current status of the database, including both the content and technique. First, we will incorporate gene family information to differentiate paralogs into different subcategories by estimating the timing of speciation and duplication. Second, we would like to develop intelligent modules to identify specific events for gene orientation and sequence changes to cope with user demands. Third, we also plan to tag evolutionarily conserved gene sets to their functional roles in terms of metabolic pathways and networks for studying mechanisms of co-regulation. Fourth, we will attempt to improve visual effects and make better gene alignment by introducing the concept of 'gaps', adding user-friendly operational options. Last, we will continue to update the database with newly acquired genomes and annotations and build automatic protocols for processing data and generating results at lesser key strikes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [31301030]; CAS Youth Innovation Promotion Association [to D.W.]. Funding for open access charge: National Natural Science Foundation of China [31301030] [to D.W.].
Conflict of interest statement. None declared.

REFERENCES

1. Elghabi,Z., Ruf,S. and Bock,R. (2011) Biolistic co-transformation of the nuclear and plastid genomes. *Plant J.*, **67**, 941–948.
2. Svab,Z., Hajdukiewicz,P. and Maliga,P. (1990) Stable transformation of plastids in higher plants. *Proc. Natl Acad. Sci. U.S.A.*, **87**, 8526–8530.

3. Gould,S.B., Waller,R.F. and McFadden,G.I. (2008) Plastid evolution. *Annu. Rev. Plant Biol.*, **59**, 491–517.
4. Bendich,A.J. (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays*, **6**, 279–282.
5. Archibald,J.M. (2009) The puzzle of plastid evolution. *Curr. Biol.*, **19**, R81–R88.
6. Smith,D.R. and Lee,R.W. (2014) A plastid without a genome: evidence from the nonphotosynthetic green algal genus Polytomella. *Plant Physiol.*, **164**, 1812–1819.
7. Molina,J., Hazzouri,K.M., Nickrent,D., Geisler,M., Meyer,R.S., Pentony,M.M., Flowers,J.M., Pelser,P., Barcelona,J., Inovejas,S.A. *et al.* (2014) Possible loss of the chloroplast genome in the parasitic flowering plant Rafflesia lagascae (Rafflesiaceae). *Mol. Biol. Evol.*, **31**, 793–803.
8. Barbrook,A.C., Howe,C.J. and Purton,S. (2006) Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci.*, **11**, 101–108.
9. Oudot-Le Secq,M.P., Grimwood,J., Shapiro,H., Armbrust,E.V., Bowler,C. and Green,B.R. (2007) Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana: comparison with other plastid genomes of the red lineage. *Molecular Genet. Genomics*, **277**, 427–439.
10. Turmel,M., Otis,C. and Lemieux,C. (1999) The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: insights into the architecture of ancestral chloroplast genomes. *Proc. Natl Acad. Sci. U.S.A.*, **96**, 10248–10253.
11. Funk,H.T., Berg,S., Krupinska,K., Maier,U.G. and Krause,K. (2007) Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, Cuscuta reflexa and Cuscuta gronovii. *BMC Plant Biol.*, **7**, 45.
12. de Koning,A.P. and Keeling,P.J. (2006) The complete plastid genome sequence of the parasitic green alga Helicosporidium sp. is highly reduced and structured. *BMC Biol.*, **4**, 12.
13. Cai,Z., Penaflor,C., Kuehl,J.V., Leebens-Mack,J., Carlson,J.E., dePamphilis,C.W., Boore,J.L. and Jansen,R.K. (2006) Complete plastid genome sequences of Drimys, Liriodendron, and Piper: implications for the phylogenetic relationships of magnoliids. *BMC Evol. Biol.*, **6**, 77.
14. Janouskovec,J., Liu,S.L., Martone,P.T., Carre,W., Leblanc,C., Collen,J. and Keeling,P.J. (2013) Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PloS One*, **8**, e59001.
15. Deusch,O., Landan,G., Roettger,M., Gruenheit,N., Kowallik,K.V., Allen,J.F., Martin,W. and Dagan,T. (2008) Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol. Biol. Evol.*, **25**, 748–761.
16. Wang,D., Wu,Y.W., Shih,A.C., Wu,C.S., Wang,Y.N. and Chaw,S.M. (2007) Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. *Mol. Biol. Evol.*, **24**, 2040–2048.
17. Race,H.L., Herrmann,R.G. and Martin,W. (1999) Why have organelles retained genomes? *Trends Genet.*, **15**, 364–370.
18. Raven,J.A. and Allen,J.F. (2003) Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biol.*, **4**, 209.
19. Davila Lopez,M., Martinez Guerra,J.J. and Samuelsson,T. (2010) Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PloS One*, **5**, e10654.
20. Ballouz,S., Francis,A.R., Lan,R. and Tanaka,M.M. (2010) Conditions for the evolution of gene clusters in bacterial genomes. *PLoS Comput. Biol.*, **6**, e1000672.
21. Liu,X. and Han,B. (2009) Evolutionary conservation of neighbouring gene pairs in plants. *Gene*, **437**, 71–79.
22. Xie,B., Wang,D., Duan,Y., Yu,J. and Lei,H. (2013) Functional networking of human divergently paired genes (DPGs). *PloS One*, **8**, e78896.
23. Stoebe,B. and Kowallik,K.V. (1999) Gene-cluster analysis in chloroplast genomics. *Trends Genet.*, **15**, 344–347.
24. O'Brien,E.A., Zhang,Y., Wang,E., Marie,V., Badejoko,W., Lang,B.F. and Burger,G. (2009) GOBASE: an organelle genome database. *Nucleic Acids Res.*, **37**, D946–D950.
25. Cui,L., Veeraraghavan,N., Richter,A., Wall,K., Jansen,R.K., Leebens-Mack,J., Makalowska,I. and dePamphilis,C.W. (2006) ChloroplastDB: the chloroplast genome database. *Nucleic Acids Res.*, **34**, D692–D696.
26. Kurihara,K. and Kunisawa,T. (2004) A gene order database of plastid genomes. *Data Sci. J.*, **3**, 60–79.
27. Maguire,S.L., OhEigeartaigh,S.S., Byrne,K.P., Schroder,M.S., O'Gaora,P., Wolfe,K.H. and Butler,G. (2013) Comparative genome analysis and gene finding in Candida species using CGOB. *Mol. Biol. Evol.*, **30**, 1281–1291.
28. Lopez,M.D. and Samuelsson,T. (2011) eGOB: eukaryotic gene order browser. *Bioinformatics*, **27**, 1150–1151.
29. Louis,A., Muffato,M. and Roest Crollius,H. (2013) Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Res.*, **41**, D700–D705.
30. Wang,D., Zhang,Y., Fan,Z., Liu,G. and Yu,J. (2012) LCGbase: a comprehensive database for lineage-based co-regulated genes. *Evol. Bioinform. Online*, **8**, 39–46.
31. Proost,S., Van Bel,M., Sterck,L., Billiau,K., Van Parys,T., Van de Peer,Y. and Vandepoele,K. (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.
32. Kandimalla,E.R., Bhagat,L., Wang,D., Yu,D., Sullivan,T., La Monica,N. and Agrawal,S. (2013) Design, synthesis and biological evaluation of novel antagonist compounds of Toll-like receptors 7, 8 and 9. *Nucleic Acids Res.*, **41**, 3947–3961.
33. Burger,G., Lavrov,D.V., Forget,L. and Lang,B.F. (2007) Sequencing complete mitochondrial and plastid genomes. *Nat. Protoc.*, **2**, 603–614.
34. Lohse,M., Drechsel,O., Kahlau,S. and Bock,R. (2013) OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.*, **41**, W575–W581.
35. Grant,J.R., Arantes,A.S. and Stothard,P. (2012) Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics*, **13**, 202.
36. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
37. Bussotti,G., Raineri,E., Erb,I., Zytnicki,M., Wilm,A., Beaudoing,E., Bucher,P. and Notredame,C. (2011) BlastR–fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.*, **39**, 6886–6895.
38. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
39. Xu,Z. and Hao,B. (2009) CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res.*, **37**, W174–W178.
40. Soltis,D.E. and Soltis,P.S. (2004) Amborella not a 'basal angiosperm'? Not so fast. *Am. J. Bot.*, **91**, 997–1001.
41. Jansen,R.K., Cai,Z., Raubeson,L.A., Daniell,H., Depamphilis,C.W., Leebens-Mack,J., Muller,K.F., Guisinger-Bellian,M., Haberle,R.C., Hansen,A.K. *et al.* (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl Acad. Sci. U.S.A.*, **104**, 19369–19374.
42. De Las Rivas,J., Lozano,J.J. and Ortiz,A.R. (2002) Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res.*, **12**, 567–583.
43. Yoon,H.S., Hackett,J.D., Ciniglia,C., Pinto,G. and Bhattacharya,D. (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.*, **21**, 809–818.
44. Yagi,Y. and Shiina,T. (2014) Recent advances in the study of chloroplast gene expression and its evolution. *Front. Plant Sci.*, **5**, 61.